An On-line Speaker Adaptation Method for HMM-based ASRs

András Bánhalmi, Dénes Paczolay, and András Kocsor

When building a robust continous speech recognition system, one good way of improving the recognition accuracy [4] is via speaker adaptation [2]. All the state-of-the-art dictation systems use techniques to adapt the initial speaker independent model to a new speaker. These techniques require a large amount of sentences for utterance by the user, and the computation of the adaptation begins after recording these predefined sentences. The dictation system allows speech recognition only after this long-time adaptation procedure has been completed.

Our goal is to avoid this time-consuming procedure by creating a method that is embedded into a continous speech recogniser to retrieve the data for the adaptation process. The method proposed by us can be used with HMM-based speech recognisers [1] where multistack structure is used to store the hypotheses. We call a "hypothesis" a phoneme series, and it has a probability at a specific time. The hypotheses are stored in a stack ordered according to their probability. When a hypothesis is being extended (the next sound frame is being evaluated), the grammar module is asked for the possible phoneme continuatations and probabilities for them. The acoustic probablilty given by the HMM and the probability given by the grammar module will change the order of the hypotheses in the next stack. If two or more hypotheses have the same phoneme series, they will be fused into one hypothesis. This can be done without losing information in HMM-based systems. Then, using Viterbi and N-best cutting we drop the least likely hypotheses from the next stack. More details can be found about systems like ours in [3].

Of course embedding an automatic adaptation data retrieval method into a system like our, will introduce some problems. Not only the phoneme series and the data of the last phoneme HMMare needed to be stored by the system, but the tables of the most probable previous state, the tables of the gaussian components, and the tables containing the information of the jumps to the next phoneme HMM's start state are also required. Moreover, to create fast recognition, it is important to decrease the numbers of the hypotheses by fusing them. If only recognition is performed, all the hypotheses with the same phoneme series can be fused, but when adaptation information is stored, not all these hypotheses can be. In this paper we propose a method for storing the adaptation data efficiently in a graph structure, then we can examine which hypotheses can be fused and which can not.

Using an adaptation data retrieval method like this, the sentences for utterances need not be predefined, as data retrieval could be run during the recognition phase. Using this technique in practice gives rise to certain other problems. The retrieved data can be used for adaptation only if they are accurate, so the user should approve the correctness of the recognition result. If the recognition result is approved, then the accuracy of the retrieved data will be high only if the segmentation borders between phonemes are well determined. Here we will compare the accuracy of our continous automatic segmentation with that automatic segmentation which is based on a predefined sequence and dynamic programming.We will use the manually segmented data as a baseline. The above mentioned automatic segmenting algorithm using a predefined phoneme series builds one large HMM model, which is a chain of the HMMs labeled by the phonemes of the sentence. When segmenting phonetically the given utterance, this large HMM model will be evaluated via the Viterbi algorithm, and the most probable path of HMM states will be traced back to give us the phonetic segmentation. The effect of these segmentation methods will be compared by the accuracy of the recognition after adaptation. Among the widely-used adaptation methods we use the well-known MLLR adaptation algorithm for the evaluation.

References

[1] C. Becchetti and L. P. Ricotti. Speech Recognition, *John Wiley and Son*, England, 2000.

[2] X. Huang, A. Acero, and H. Hon. Spoken Language Processing, *Englewood Cliffs, NJ, Prentice Hall, 2001*.

[3] A. Kocsor, A. Bánhalmi, D. Paczolay, J. Csirik, and L. Pávics. The Oasis Speech Recognition System for Dictating Medical Reports, *Annual Congress of the European Association of Nuclear Medicine, 05, 2005*.

[4] L. Neumeyer, A. Sankar, and V. Digalakis. A Comparative Study of Speaker Adaptation Techniques, *Proceedings Eurospeech, pages 1127-1130, 1995.*